LSAC RESEARCH REPORT SERIES

■ A Simulation Study of Optimal Online Calibration of Testlets Using Real Data

Douglas H. Jones
Mikhail S. Nediak
Rutgers, The State University of New Jersey

■ Law School Admission Council
Computerized Testing Report 00-07
December 2005

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary 1

Abstract 1

Introduction 2

IRT Model, Optimal Design Criterion, Weighted Hellinger Deviance, and Bayes Estimation Using MCMC 3

Simulation Description and Sequential Design Schemes 4

Performance Measure 4

Results 5

Discussion 7

Concluding Remarks 7

References 8

Appendix: Hellinger Deviance 9

Executive Summary

This research investigates new ways to estimate the characteristics of pretest items, grouped by their relation to a reading passage or other common stimulus. Currently, practitioners use a method called *spiraling* to obtain a random sample of item response data for estimating item characteristics. However, a random sample of item responses cannot tell how items that are difficult for average-ability test takers would perform for high-ability test takers. In addition, random samples of item responses provide an excess of information for items of average difficulty and discrimination.

These new methods take advantage of the online test-taking environment by selecting the best test takers for the new pretest items. Thus, items receive responses from test takers whose abilities can provide the most information about the items' characteristics. Estimates of test takers' abilities are available in the online setting from their performances on operational sections of a test. Response data are taken in batches, and after each new batch of data, item characteristics of the new items are re-estimated. The process stops sampling after enough information is gathered or a limit on the sample size is reached. The new methods are based on established statistical methods in optimal design of experiments, started in the 1920's by Sir R. A. Fisher, and sequential sampling, started by Dr. A. Wald during WWII.

Up to this time, research has focused on the online collection of item-response data for several unrelated items either in isolation or simultaneously. However, no practical solutions have addressed online collection of item-response data for items in natural groupings, or item sets, such as testlets or common stimulus formats. In this paper, we will assume a situation where a calibration session must collect online data about several pretest testlets simultaneously.

We wish to compare the efficacy of the two sequential optimal design schemes and a spiraling scheme. This investigation uses 28,184 live responses to 101 LSAT items grouped into testlets by common reading-passage stimuli. Online testing is simulated with the collection of 800 responses per testlets. Twenty-three items naturally grouped into four testlets are calibrated. The remaining 78 items provide the basis for test-taker ability estimates.

For most items, our sequential optimal methods performed very well. However, two testlets, representing over half of the items, were not well estimated by the spiraling design. This is due to the extreme values of some of the item parameters. In these cases, the sequential optimal designs achieved superior performance by using information about the extreme parameter values to push the most informative test takers to these testlets.

The results imply that sequential optimal designs perform much better than spiraling designs, even when items must be given in testlets format. The cost of implementing a sequential optimal design should be made up by avoiding the opportunity loss from use of a spiraling design.

With a sequential optimal design, it is feasible that accuracy of calibration could be tolerable at sample sizes as small as 600. Such a small sample size, versus the sample size for accurate spiraling, has many implications; for instance, sequential optimal methods could calibrate many more items than spiraling in the same time period.

Abstract

Many standardized tests use formats called testlets, in which a passage or stem is presented followed by a group of related questions. This paper presents a simulation study of an online item-calibration method for the testlets format that employs sequentially optimal methods. Item calibrations are continually updated using response data sequentially aggregated over small batches of online test takers. Within batches of test takers, testlets compete for responses according to optimally derived sampling designs. Between batches of samples, item parameter estimates are recalculated and optimal designs are derived for the next batch of test takers. The derived designs optimize the determinant of the Fisher information matrix. A conjugate-gradient-search algorithm is used to optimize this nonlinear objective.

It is known that sequentially optimal methods are extremely sensitive to unstable parameter estimates, causing inefficient sample designs. To overcome estimation instability, Bayesian methods are employed to estimate item parameters via a Markov Chain Monte Carlo (MCMC) scheme especially adapted for sequential estimation.

This investigation uses 28,184 real responses to 101 LSAT items grouped by the testlets format. Online testing is simulated by random draws from this dataset. The optimal design scheme is compared to spiraling, a totally random data collection method. As expected, the results favor optimally collected data over spiraling. Computational times for optimal designs are well within bounds for implementation in an actual setting. More refined Bayes procedures would provide more stability, which would allow collection of optimal data earlier, and thus further improve performance over spiraling.

Introduction

In an online adaptive test, data for calibrating new items are collected from examinees during the test. Typically, a representative sample of the ability in the test-taking population is obtained by spiraling new items throughout the operational test. Spiraling is a form of systematic sampling, and if the test takers are randomly ordered, it produces a random sample of item-response data from the test-taker population. A large portion of a representative sample of test takers will provide good information for estimating the parameters of an item of average difficulty. However, if item parameters are extreme in either difficulty (high or low) or discrimination (high), they will be poorly estimated by a representative sample of test takers. In addition, even for an item of average difficulty, about 15% to 25% of the data in a representative sample are uninformative and thus wasted (Jones & Jin, 1994).

Optimal design theory provides a framework for obtaining samples of response data that are most informative in the criterion chosen by the investigator. Speaking in general terms, one is interested in obtaining a sample that gives the highest Fisher criterion for parameter estimates. Formulas are given in Jones and Jin (1994), and elsewhere, that prescribe the set of test takers yielding highest Fisher information for any item. Unfortunately, these formulas must know already the values of the item parameters. However, if approximate values of item parameters and ability parameters are available, then it is possible to match test takers and items to yield much more precise estimates than are possible with spiraling. Approximate values of item and ability parameters could be either point estimates or prior distributions. Approximate values of ability parameters are available in online testing. The online testing operation could be designed so that approximate values of item parameters are available from previous stages of test administration.

A solution, then, for obtaining more information than provided by spiraling is based on sequential optimal design. The basic idea is to obtain sample data in batches, where, after each batch, item parameters are recalibrated. Then these new item parameter estimates are used with an optimal design method to prescribe the next batch of test takers for each new item. Other researchers have recognized the possibility of using sequential optimal design methods to calibrate items (see van der Linden, 1988; Berger, 1991, 1992, 1994; Berger & van der Linden, 1991; Jones & Jin, 1994; Jones, Chiang, & Jin, 1997; and Jones, Nediak, & Wang, 1999).

Up to this time, research has focused on the online collection of item response data for several unrelated items, either in isolation or simultaneously, using the two- or three-parameter logistic model. This research is theoretically interesting since it uses the optimal design framework and the solution methods involve constrained nonlinear optimization. The objective, Fisher information, is nonlinear and requires nonlinear programming for optimization, such as conjugate-gradient-search (Murtaugh & Saunders, 1978; and Jones, Nediak, & Wang, 1999). However, no practical solutions have addressed online collection of item-response data for items in natural groupings, or item sets, such as testlets or common stimulus formats. In this paper, we will assume a situation where a calibration session must collect online data about several pretest testlets simultaneously.

We propose two variants of sequential optimal design methods for the testlet-response data-collection problem. They are called the *constrained* and the *unconstrained* methods. Both methods collect data in batches and recalibrate item parameters after each batch. However, they differ in the way they obtain optimal designs for the next batch of data. The unconstrained method matches a batch of test takers with several testlets to optimize the Fisher information without any constraints in the number of test takers going to a particular testlet. The constrained method puts constraints on the number of test takers each testlet may have. The reason one may want to use the constrained method is to avoid the possibility that one testlet monopolizes all of the test takers, leaving the other testlets uncalibrated.

We wish to compare the efficacy of the two sequential optimal design schemes and a spiraling scheme. This investigation uses 28,184 live responses to 101 LSAT items grouped into testlets by common reading passage stimuli. Online testing is simulated with the collection of 800 responses per testlet.

The next section introduces the three-parameter logistic model and the Fisher information for the item parameters. The optimal design criterion is based on the determinant of the Fisher information matrix. Item parameter estimates are compared through the values of the associated item response functions (IRFs) using the weighted Hellinger deviance, which is defined in the appendix. The Bayes posterior distribution of the item parameters is used extensively, not only to estimate item parameters, but also to average the design criterion over the parameter space. The posterior distribution is obtained using Markov Chain Monte Carlo (MCMC) simulation introduced for IRT estimation in Patz and Junker (1999). Following this section, the online design schemes are described along with the simulation methods to compare their performance.

IRT Model, Optimal Design Criterion, Weighted Hellinger Deviance, and Bayes Estimation Using MCMC

This section describes theoretical selection models for sequentially optimal methods. Bayes posteriors of item parameters are continually updated using response data sequentially aggregated over small batches of online test takers. Within batches of test takers, testlets compete for responses according to optimally derived sampling designs. Between batches of samples, item parameter posteriors are recalculated and optimal designs are derived for the next batch of test takers.

Let u_i denote a response to a single item from individual i with ability level θ_i. Let β denote unknown parameters associated with a particular item. Assume that all responses are scored either correct, $u_i = 1$, or incorrect, $u_i = 0$. The item-response function $P(\theta_i ; \beta)$ describes the probability of correct response by a test taker with ability level θ_i. The mean and variance of the parametric family are

$$E\{u_i | \beta\} = P(\theta_i ; \beta) \tag{1}$$

$$\sigma^2(\theta_i ; \beta) = Var\{u_i | \beta\} = P(\theta_i ; \beta)[1 - P(\theta_i ; \beta)]. \tag{2}$$

We shall focus on the family of three-parameter logistic (3PL) response functions:

$$P(\theta_i ; \beta) = \beta_2 + (1 - \beta_2)R(\theta_i ; \beta), \tag{3}$$

where

$$R(\theta_i ; \beta) = \frac{e^{\beta_0 + \beta_1 \theta}}{1 + e^{\beta_0 + \beta_1 \theta}}. \tag{4}$$

The three-parameter item IRT characteristics are: the *discrimination power*, $a = 1.7\beta_1$; the *difficulty*, $b = -\beta_0 / \beta_1$; and *guessing*, $c = \beta_2$. The family of two-parameter logistic (2PL) response functions is obtained from expression (3) and $\beta_2 = 0$.

The expected information from a response at ability θ is defined as the Fisher information matrix. The 3PL information matrix is

$$m(\beta;\theta) = \frac{[1 - P(\theta;\beta)]^2}{P(\theta;\beta)[1 - P(\theta;\beta)]} \begin{pmatrix} (1-\beta_2)^2[R(\theta;\beta)]^2 & (1-\beta_2)^2[R\theta;\beta)]^2\theta & (1-\beta_2)R(\theta;\beta) \\ (1-\beta_2)^2[R(\theta;\beta)]^2\theta & (1-\beta_2)^2[R(\theta;\beta)]^2\theta^2 & (1-\beta_2)R(\theta;\beta)\theta \\ (1-\beta_2)R(\theta;\beta) & (1-\beta_2)R(\theta;\beta)\theta & 1 \end{pmatrix} \tag{5}$$

Suppose that we have a set T of items, which is partitioned into m nonintersecting subsets $T(j)$, $j = 1, \dots , m$ (testlets). Assume point estimates of test takers' abilities are available, denoted by $\Theta = \{\theta_i : i = 1, \dots , n\}$. Point estimates of item parameter vectors are denoted by $\mathbf{B} = \{\beta_t : t \in T\}$. Introduce x_{ij} equal to the number of observations taken for testlet j from examinee i, and let $\mathbf{x} = (x_{11}, \dots, x_{1m}; x_{21}, \dots, x_{2m}; x_{n1}, \dots, x_{nm})$. In the following, we describe the information matrix, $\mathbf{M}(\mathbf{B}; \Theta, \mathbf{x})$, for all the testlet item parameters, $\mathbf{B}$, available with the design, $\mathbf{x}$, over the abilities, Θ.

The information matrix $\mathbf{M}^{(t)}$ of item t belonging to testlet $T(j)$ is

$$\mathbf{M}^{(t)} \equiv \mathbf{M}^{(t)}(\beta_t ; \Theta_{,} \mathbf{x}) = \sum_{i=1}^{n} x_{ij} \mathbf{m}^{(i,t)}, \tag{6}$$

where $\mathbf{m}^{(i,\,t)} = \mathbf{m}(\beta_t ; \theta_i)$. The information matrix $\mathbf{M}(\mathbf{B}; \Theta, \mathbf{x})$ of all the items in a testlet is block diagonal with respect to individual items. Thus, the design criterion used is

$$\log \det \mathbf{M}(\mathbf{B}; \Theta, \mathbf{x}) = \sum_{j=1}^{m} \sum_{t \in T(j)} \log \det \mathbf{M}^{(t)}(\beta_t ; \Theta , \mathbf{x}). \tag{7}$$

A design **x** is called *exact* if x_{ij} is an integer for all i's and j's; otherwise, it is called *approximate*. An optimal design will typically be an approximate design. The design can be constrained by bounds on the number of test takers going to testlets. We will study both constrained and unconstrained versions. The design specifies the probability of observing certain parings between test takers and testlets. Optimal designs are derived using a projected-gradient-search algorithm especially formulated for this problem (see Jones, Nediak, & Wang, 1999).

Empirical Bayes methods are used to estimate item parameters. Bayes prior distributions are the posteriors from the previous sampling stages. Posterior estimation employs Markov Chain Monte Carlo (MCMC) simulation based on Patz and Junker (1999) and Jones and Nediak (2000). To increase the MCMC acceptance rate, samples from the previous stage are used for the proposal sample in the current stage. To increase stability further, the Fisher information matrix is replaced in the optimization criterion by its average over the posterior.

Simulation Description and Sequential Design Schemes

The simulation creates online samples from a set of real data obtained in a paper-and-pencil administration of the LSAT. The sample space consists of the responses to 101 items, grouped into testlets by common reading passage stimuli, from 28,184 test takers. The 23 items coming from the first four testlets, of size 6, 5, 6, and 6 items, are chosen to calibrate. This leaves 78 items to estimate ability parameters.

These ability estimates are used as if the test takers were in the process of taking an online test with their abilities having been estimated from a portion of the test. Simulated online item calibration uses these data by randomly choosing test-taker records and their actual responses to a particular item.

The optimal design process begins by choosing a random batch of test-taker data. Their ability parameters and a posterior distribution of item parameters are fed to the mathematical program, which derives the design. The design specifies which test-taker data are used to estimate the parameters of each testlet. Using the updated posterior distribution of item parameters after each batch of data, this process continues for several batches of test takers until the limiting sample size is reached.

The limiting sample size is 80 batches of test takers multiplied by 40, yielding 3,200 total responses. This sample is spread among the four testlets in two ways. The first way, called the constrained case, requires each batch of 40 test takers to yield 10 new observations for each testlet. The second way, called the unconstrained case, makes no such requirement, and it is possible that one testlet can monopolize the whole batch of test takers.

The posterior distribution of item parameters is generated with responses from 40 randomly selected batches of 40 test takers. Using this posterior, optimal assignment of 40 test takers is generated. This step is repeated for each batch of test takers to update the posterior distribution until, finally, 40 batches of test takers have been assigned optimally. Thus, 1,600 optimal responses and 1,600 totally random responses are obtained for 3,200 responses. This results in 800 observations per testlet for the constrained designs and approximately 800 observations (hopefully) for the unconstrained designs. Summaries of the designs follow.

Unconstrained Design: Sequential estimation using a posterior distribution formed from 40 batches of size 40 randomly selected test takers, followed by 40 batches of size 40 optimally selected test takers. The optimal design scheme allows more observations for testlets that require more information. For testlets with roughly the same mix of item parameters, this results in approximately 800 responses per testlet.

Constrained Design: Sequential estimation using a posterior distribution formed from 40 batches of size 40 randomly selected test takers, followed by 40 batches of size 40 optimally selected test takers. The optimal design scheme gives equal numbers of observations for testlets. This results in exactly 800 responses per testlet.

Spiraling Design: No sequential estimation, 80 batches of size 40 randomly selected test takers. This results in exactly 800 responses per testlet.

Performance Measure

The closeness, or lack of closeness, between estimated and target parameter values does not clearly measure how the procedures performed. Values of the item parameters compensate each other to fit the IRF. Instead of comparing item parameters directly, we will measure the distance between the estimated and target IRFs evaluated at particular ability values. This study uses target IRFs estimated from all 28,184 responses. The Hellinger deviance, defined in the Appendix, is used to measure the fit between values of the estimated and target IRFs. The Hellinger deviance is a metric for a set of probability distributions forming points in an abstract space. We can construct a value that measures deviations over an ability range by weighting the Hellinger deviance with respect to a distribution over ability. We chose *N*(1, 1) for this distribution, since we wanted to see which designs could satisfy tight tolerance bounds between estimated and target IRFs in the range zero to two. The ability scale of the real data has been transformed so that the ability distribution is *N*(0, 1). This distribution and the distribution for the weighted Hellinger deviance are chosen independently of each other.

Note that the comparison between estimated and target IRFs measures the bias in estimation. Although our optimality criterion is based on asymptotic variance measures, the sample sizes are so small that bias is the overriding concern for item parameter estimation.

Results

The first 25 lines of data of a typical calibration sample, for either optimal design or the spiraling design, are displayed in Table 1. As usual, 0 and 1 indicate incorrect and correct responses. A 3 indicates that the item was not administered to the test taker. Our custom item-calibration program ignores responses coded with 3s. In addition, typical calibration programs, such as BILOG or LOGIST, could be used with this data-coding scheme to calibrate these items, since these programs treat a 3 as "not reached."

TABLE 1
Generic online calibration sample showing 25 test takers (0, incorrect; 1, correct; 3, not administered)

	Testlet																						
	1						2					3						4					
Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	1	1	1	0	1	1
2	3	3	3	3	3	3	0	0	1	0	1	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	0	1	0	0	1	3	3	3	3	3	3	3	3	3	3	3	3
4	3	3	3	3	3	3	0	0	1	0	1	3	3	3	3	3	3	3	3	3	3	3	3
5	3	3	3	3	3	3	0	0	1	0	1	3	3	3	3	3	3	3	3	3	3	3	3
6	1	1	0	1	1	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
7	3	3	3	3	3	3	0	0	1	1	0	3	3	3	3	3	3	3	3	3	3	3	3
8	3	3	3	3	3	3	3	3	3	3	3	0	1	1	0	1	0	3	3	3	3	3	3
9	3	3	3	3	3	3	0	0	0	0	1	3	3	3	3	3	3	3	3	3	3	3	3
10	1	1	0	0	1	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
11	0	0	0	0	1	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
12	3	3	3	3	3	3	3	3	3	3	3	1	0	1	1	0	0	3	3	3	3	3	3
13	3	3	3	3	3	3	3	3	3	3	3	1	0	1	1	0	0	3	3	3	3	3	3
14	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	1	1	1	1	0	1
15	3	3	3	3	3	3	0	0	0	0	0	3	3	3	3	3	3	3	3	3	3	3	3
16	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	0	0	1	1	1	0
17	0	1	1	0	1	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
18	3	3	3	3	3	3	0	1	0	1	0	3	3	3	3	3	3	3	3	3	3	3	3
19	0	0	1	1	1	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
20	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	1	0	1	1	0	1
21	1	1	1	0	1	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
22	3	3	3	3	3	3	3	3	3	3	3	1	1	1	0	1	1	3	3	3	3	3	3
23	1	0	1	1	0	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
24	3	3	3	3	3	3	3	3	3	3	3	0	1	1	1	0	0	3	3	3	3	3	3
25	3	3	3	3	3	3	3	3	3	3	3	1	1	0	1	0	1	3	3	3	3	3	3

The target and estimated item parameters are displayed in Table 2. The weighted Hellinger deviance (in the column denoted by H in Table 2) has been multiplied by 10,000. The final item and ability parameter estimates were put on a common scale.

TABLE 2
Target and sequential MCMC parameter estimates

		Target IRF			Constrained Sequential				Unconstrained Sequential					Spiraling			
Testlet	ID	b_0	b_1	b_2	b_0	b_1	b_2	H	n	b_0	b_1	b_2	H	b_0	b_1	b_2	H
1	1	0.17	0.80	0.13	–0.02	0.90	0.18	1	777	0.09	0.82	0.17	0	–0.33	0.98	0.32	3
	2	0.63	0.96	0.3	1.05	1. 15	0.18	29	777	0.87	1.25	0.23	28	1.01	1.20	0.18	30
	3	–0.24	1.17	0.05	–0.20	1.32	0.03	10	777	–0.15	1.25	0.02	5	–0.16	1.31	0.04	14
	4	–1.05	1.18	0.14	–1.11	1.31	0.11	7	777	–0.89	1.16	0.09	2	–0.85	1.19	0.07	5
	5	0.44	0.70	0.14	0.54	0.78	0.16	14	777	0.24	0.93	0.24	17	0.42	0.77	0.23	13
	6	1.16	0.57	0.17	0.90	0.68	0.23	5	777	0.80	0.72	0.27	5	0.87	0.71	0.29	3
2	7	1.28	0.68	0.00	0.79	0.71	0.23	5	743	0.70	0.77	0.29	2	0.78	0.78	0.21	7
	8	–1.62	1.27	0.11	–2.06	1.54	0.12	11	743	–1.89	1.54	0.12	13	–1.42	0.97	0.09	33
	9	0.82	0.84	0.02	0.45	0.82	0.19	7	743	0.19	0.88	0.24	20	0.39	0.81	0.19	12
	10	–1.29	1.16	0.24	–0.84	0.91	0.12	12	743	–1.23	1.17	0.18	8	–1.36	1.06	0.24	13
	11	0.57	0.73	0.22	0.24	0.72	0.36	2	743	0.14	0.78	0.34	5	0.19	0.98	0.34	8
3	12	–2.27	1.31	0.17	–2.88	1.69	0.19	10	877	–2.91	1.68	0.19	9	–3.32	1.89	0.21	17
	13	–1.02	1.01	0.24	–1.28	1.22	0.24	7	877	–1.48	1.33	0.25	13	–1.45	1.33	0.27	11
	14	–1.96	1.05	0.27	–2.98	1.53	0.34	9	877	–3.67	1.81	0.35	17	–3.98	2.04	0.35	28
	15	–2.24	1.10	0.14	–2.13	1.10	0.12	1	877	–2.07	1.01	0.12	1	–1.93	0.88	0.09	16
	16	–3.45	1.43	0.10	–4.04	1.66	0.09	10	877	–3.98	1.78	0.10	11	–4.73	1.89	0.11	19
	17	–1.54	0.91	0.20	–2.09	1.15	0.23	8	877	–2.34	1.26	0.24	15	–2.19	1.15	0.22	21
4	18	0.27	0.56	0.01	–0.27	0.83	0.17	13	803	–0.30	0.77	0.20	5	–0.16	0.82	0.18	16
	19	–0.82	0.89	0.12	–0.93	0.97	0.17	6	803	–0.75	0.83	0.11	1	–0.71	0.86	0.07	1
	20	–1.96	0.99	0.10	–1.77	0.97	0.07	2	803	–1.69	0.82	0.06	8	–1.92	1.02	0.08	2
	21	–3.00	1.18	0.09	–2.67	1.04	0.08	2	803	–2.99	1.09	0.10	5	–3.01	1.23	0.09	1
	22	–2.45	0.99	0.14	–1.95	0.80	0.10	2	803	–2.00	0.78	0.08	7	–2.14	0.80	0.13	5
	23	–2.07	0.90	0.24	–3.81	1.61	0.30	18	803	–3.37	1.44	0.30	11	–2.99	1.16	0.28	13

Due to the compensatory effect in estimating item parameters, with such small samples, it is unsurprising that parameter estimates are not close to their target values. However, without drawing all the associated IRF curves, the weighted Hellinger deviancies give a clear and simple picture of how close estimated and target IRFs are. A deviance value above 15 indicates there is serious departure from the target IRF. Note in particular that testlets 2 and 3 are not well estimated by the spiraling design. This is due to the extreme values of the item parameters. In these cases, the sequential optimal designs have used information about the extreme parameter values to push the most informative test takers to these testlets. In general, sequential optimal designs generally obtained smaller deviancies than spiraling designs.

However, even sequential designs can perform poorly, as indicated by item 2. The poor performance is due to the fact that item 2 is extreme among its item set. No form of optimal design methods could overcome such a configuration.

Figure 1 displays cumulative distributions of the weighted Hellinger deviancies for all three procedures. Approximately 70% of the IRFs estimated by constrained sequential designs are acceptably close to their targets as compared to 40% estimated by the spiraled designs using a deviance value of 10. In addition, 91% of the IRFs estimated by constrained sequential designs are acceptably close to their targets as compared to 65% estimated by spiraled designs using a deviance of 15. The performance of the constrained and the unconstrained sequential procedures are approximately the same, with a slight edge going to the constrained method.

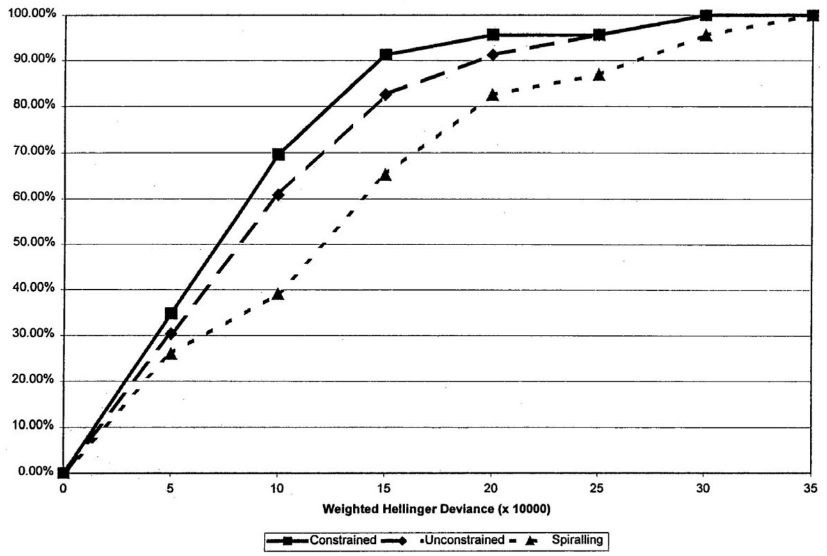

FIGURE 1. *Cumulative distribution of weighted Hellinger deviance comparing constrained optimal, unconstrained optimal, and spiraling designs*

Discussion

The results imply that sequential optimal designs perform much better than spiraling designs, even when items must be given in testlet format. The cost of implementing a sequential optimal design should be made up in avoiding the opportunity loss from using a spiraling design.

As we have seen, the spiraling design frequently gives an inferior calibration result for some extreme items. More insidiously, practitioners would never know that these item parameter estimates were worthless. Note that the Fisher information or the shape of the likelihood function would not help the practitioner in this case due to the small sample sizes and the overriding bias. The only way to overcome this problem is to build redundancies into the calibration system or, in effect, take a much larger sample. However, the cost of redundancies could easily outweigh the cost of a system using sequential optimal methods.

With a sequential optimal design, it is feasible that accuracy of calibration could be tolerable at sample sizes as small as 600. Such a small sample size, versus the sample size for accurate spiraling, has many implications, one of which is that sequential optimal methods could calibrate many more items than spiraling in the same time period.

Concluding Remarks

Sequential estimation with D-optimal designs improves testlet calibration over the spiraling design considered in this paper. Clearly, some items were calibrated more effectively than others. In actual implementation of online item calibration, the more easily calibrated testlets could be removed from sampling earlier.

The estimation would be better with informative priors. Linear hierarchical priors may be employed to obtain priors that are more informative. These may be obtained from past calibrations and, possibly, information about the item. Other design criteria could be employed using stopping rules based on the variation measures in the posterior of the item parameters. Computational times for deriving optimal designs are well within bounds for implementation in an actual setting. Batch sizes and the number of testlets could be increased. In a related study, the researchers have derived designs for 27 testlets for any size batch.

References

Berger, M. P. F. (1991). On the efficiency of IRT models when applied to different sampling designs. *Applied Psychological Measurement, 15*, 283–306.

Berger, M. P. F. (1992). Sequential sampling designs for the two-parameter item response theory model. *Psychometrika, 57*, 521–538.

Berger, M. P. F. (1994). D-Optimal sequential designs for item response theory models. *Journal of Educational Statistics, 19*, 43–56.

Berger, M. P. F., & van der Linden, W. J. (1991). Optimality of sampling designs in item response theory models. In M. Wilson (Ed.), *Objective measurement: Theory into practice*. Norwood, NJ: Ablex Publishing.

Jones, D. H., & Jin, Z. (1994). Optimal sequential designs for online item estimation. *Psychometrika, 59*, 57–75.

Jones, D. H., Chiang, J., & Jin, Z. (1997). Optimal designs for simultaneous item estimation. *Nonlinear Analysis, Methods & Applications, 30(7)*, 4051–4058.

Jones, D. H., & Nediak, M. (2000). *Item parameter calibration of LSAT items using MCMC approximation of Bayes posterior distributions* (RUTCOR Research Report 7-2000). Rutgers, The State University of New Jersey.

Jones, D. H., Nediak, M., & Wang, X. (1999). *Sequential optimal designs for online item calibration* (RUTCOR Research Report 2-99). Rutgers, The State University of New Jersey.

Murtaugh, B., & Saunders, M. (1978). Large scale linearly constrained optimization. *Math Program, 14*, 42–72.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146–178.

Prakasa Rao, B. L. S. (1987). *Asymptotic theory of statistical inference*. New York: Wiley.

van der Linden, W. J. (1988). *Optimizing incomplete sampling designs for item response model parameters* (Research Report No. 88-5). Enschede, The Netherlands: University of Twente.

Appendix

Hellinger Deviance

The Hellinger deviance between two discrete probability distributions $f_1(z)$ and $f_2(z)$ is:

$$H = \sum_{i=1}^{\infty} [f_1^{1/2}(z_i) - f_2^{1/2}(z_i)]^2$$

(Prakasa Rao, 1987). The maximum value of the Hellinger deviance is two, and the minimum is zero.

If β^0 and β^1 parameter values of two IRFs, the Hellinger deviance between IRFs at a particular ability level θ is (after a simple transformation):

$$H(\beta^0, \beta^1; \theta) = 2\Big[1 - \sqrt{P(\theta; \beta^0)P(\theta; \beta^1)} - \sqrt{[1 - P(\theta; \beta^0)][1 - P(\theta; \beta^1)]}\Big].$$

The weighted-integral Hellinger deviance between IRFs is obtained by integrating out θ with respect to some suitable density function $w(\theta)$:

$$H(\beta^0, \beta^1) = 2\int_{-\infty}^{+\infty} H(\beta^0, \beta^1; \theta) w(\theta) d\theta.$$